Review and Analysis of some Metamodeling Techniques used in Optimization

Abderrahmane Benzaoui, Régis Duvigneau

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Review and Analysis of some Metamodeling Techniques used in Optimization

Abderrahmane BENZAOUI — Regis DUVIGNEAU

N° 6973

June 2009

Thème NUM

ISSN 0249-6399 ISRN INRIA/RR--6973--FR+ENG

Review and Analysis of some Metamodeling Techniques used in Optimization

Abderrahmane BENZAOUI , Regis DUVIGNEAU

Thème NUM — Systèmes numériques
Projet OPALE

Rapport de recherche n° 6973 — June 2009 — 17 pages

Abstract: This papers aims at presenting a review of some metamodels used in optimization. We are interested in particular in the Radial Basis Functions and Kriging metamodels. The theory of these techniques is presented and their implementation is discussed. We present also the differentiation of these metamodels. Furthermore, a slight comparison between Radial Basis Functions and Kriging metamodels is given in ordrer to show their similarities and differences.

Key-words: Optimization, Metamodels, Radial Bais Function, Kriging.

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles, BP 93, 06902 Sophia Antipolis Cedex (France)
Téléphone : +33 4 92 38 77 77 — Télécopie : +33 4 92 38 77 65

Revue et Analyse de certains Metamodèles utilisés en Optimisation

Résumé : Ce rapport a pour objectif de présenter une revue de quelques métamodèles utilisés en optimisation. On s'intéresse plus particulièrement aux métamodèles de type Fonctions à Bases Radiales et de type Krigeage. La théorie de ces méthodes est présentée et leur implémentation est discutée. On présente également la différentiation de ces métamodèles. Par ailleurs, une comparaison succincte entre les Fonctions à Bases Radiales et le Krigeage est donnée afin de montrer leur similarités et leur différences

Mots-clés : Optimisation, Métamodèles, Fonctions à Bases Radiales, Krigeage.

1 Introduction

In optimization problems, as in many other engineering applications, the objective function needs to be called many times, which results in a high computational cost. Then, to obtain a feasible expense, it is necessary to reduce the cost of the evaluation of this function. This reduction can be obtained by the use of surrogate models instead of the exact ones. The surrogate models are used only to accelerate the optimization procedure, but the exact evaluation is necessary, at least at the end of the procedure, to get an accurate solution. The surrogate models can be: a simplified model of the physical problem, a less accurate numerical resolution of the equations involved in the problem, or a metamodel based on the fitting of some available data [2, 5, 6].

Metamodels are widely used in optimization. They interpolate or approximate a given function by a representation of polynomials, exponentials or any other basis functions. The interpolation is carried out on a given set of point (observation points) where the function values are known. There are many types of metamodels which can be distinguished by the choice of the basis functions or by the way the interpolation is carried out. Jones [7] gives a taxonomy of metamodels used in optimization problems and the way they are integrated in the computational algorithms. In this paper, we are interested only in Radial Basis Functions (RBF) and Kriging metamodels which are widely used in optimization procedures. In RBF metamodels, the predicted value is obtained by means of some functions of the distance between the predicted point and the observation ones. However, in kriging metamodels, the prediction is made using some statistical considerations on the observation points.

In this article, we present briefly the theory of these techniques and discuss their efficiency. The differentiation of these metamodels is also presented. Kriging theory is particularly detailed and introduced with two different points of view. Indeed, in the litterature, one can find different presentations of this theory, where the connexion between them is not straightforward. After this presentation, a comparison is made between the RBF and the Kriging metamodels in order to shows the similarities and the differences between these techniques. We discuss also the difficulties generally encountered to implement this modeling tools. Finally, the article is terminated by a conclusion summarizing the main reached points.

2 Definition and utility of Metamodels

A metamodel is a "model of a model". It is a way to interpolate a given function f by a representation of polynomials, exponentials or any other basis functions. In some complex design problems, one can be interested by only a "low-cost" estimation, not necessarily exact, of the function f for some design variables values. A metamodel is used to simplify the complex underlying phenomena of this problem by considering it as a black box for which the input are the design variables and the output is a predicted value of the function f. For instance, in aerodynamic shape optimization, one can be interested in quickly computing the drag coefficient C_D for a given design vector X. Of course, the drag is a result of a

complex flow around a wing whose shape is a function of the design vector X. The exact evaluation of the drag coefficient requires the computation of the air flow using an adequate CFD solver which can be very expensive. Since the optimization algorithm needs to test a large number of the design vector values, it is very useful to find a simple relationship between the drag coefficient and the design vector.

A metamodel is not derived from the equations governing the complex phenomenon but is constructed by an empirical method. Given a set of N observations $(X_i,\ f_i = f(X_i))$, the unkown value of f at the point X is predicted by combining the known values f_i with some basis functions Φ_i using a set of additional parameters. There are many types of metamodels which can be distinguished by the choice of the basis functions and the way the additional parameters are evaluated. In this paper, we will not give an exhaustive list of the metamodel techniques, but we will present only Radial Basis Functions and Kriging metamodels.

3 Radial Basis Functions metamodels (RBF)

3.1 Presentation

In this technique, which can be seen also as an artificial neural network, the values of the function are known in N points X_i called the centers of the metamodel. The radial basis functions Φ_i are not a function of the design vector X itself, but of the distance, in the sense of certain norm (the Euclidean norm in our study), between the point X and the centers X_i:

$$\Phi_i(X) = \Phi(\parallel X - X_i \parallel) \tag{1}$$

where $\parallel X - X_i \parallel$ is the distance between X and X_i. In general, the values of Φ_i tend to zero when the point X is very far from the center X_i. The predicted value $\tilde{f}(X)$ is then a combination of the basis functions using some weights ω_i:

$$\tilde{f}(X) = \sum_{i=1}^{N} \omega_i \Phi(\parallel X - X_i \parallel) \tag{2}$$

The weights are calculated in such a way that the predicted function exactly reproduces the set of observations, i.e we must have $\tilde{f}(X_i) = f(X_i)$ for each center X_i. This leads to the linear system:

$$A \cdot W = F \tag{3}$$

where $A = (\Phi(\parallel X_i - X_j \parallel))_{i=1,N,j=1,N}$, $W = (\omega_1, ..., \omega_N)^T$ and $F = (f_1, ..., f_N)^T$.

The matrix A is symmetric and can be or not unconditionally positive definite according to the choice of the radial basis function [4]. There are several possible choices of the radial basis functions. Chandrashekarappa and Duvigneau [4] give some exemples of this functions. In this paper, we use a gaussian basis function which can be written as:

$$\Phi(r) = e^{-r^2/a_f^2} \tag{4}$$

where a_f is an attenuation factor. This function is known to give an unconditionally positive definite matrix A.

The choice of the attenuation factor is very important in the construction of the metamodel. This phase, called the training phase, is the most CPU consumer in the RBF network construction, but is a determining phase. A bad attenuation factor may lead to a bad prediction of the function f or a high condition number of the matrix A, but this depends on the set of the observed points and on the behaviour of the function itself. Many techniques were proposed to well choose a value of this factor. See for instance [4] and [11]. For instance, one can choose the Rippa technique [11], which is a leave-one-out one. A very good presentation of this technique can be found in [4]. The best attenuation factor is thus the factor that gives the smallest error in the sense of the leave-one-out technique. This is therefore an optimization problem, known to be multimodal. Because of its multimodality, this problem can be solved using the Particle Swarm Optimization method (PSO) for instance, which is more able to avoid local minima.

3.2 Differentiation of the RBF metamodel

One of the advantages of the RBF metamodel, is not only that it provides us a good approximation of the desired function, but also it can predict the gradient, $grad(f)$, and the Hessian $H(f)$ of this function with a very low cost. Indeed, without metamodel and using a finite-difference discretisation for instance, the computation of $H(f)$ requires $O(n^2)$ evaluations of the function f, where n is the dimension of the design vector X. Each evaluation of the function f can be very expensive. Using the RBF metamodel, as we shall see, one can compute the Hessian matrix or the gradient by simply differentiating the equation (2), once the metamodel has been constructed.

Let $r_l = \| X - X_l \|$ for $l = 1, N$, $K = (\Phi(r_1), ..., \Phi(r_N))^T$, and $\delta X = (X - X_1, ..., X - X_N)$. Thus, the equation (2) can be rewritten as:

$$\tilde{f}(X) = K^T \cdot W \tag{5}$$

Then, the derivative of the predicted function with respect to the i^{th} component of X is:

$$grad(\tilde{f}(X))_i = \frac{\partial \tilde{f}(X)}{\partial x_i} = Ki^T \cdot W \tag{6}$$

where $Ki = (\frac{\partial \Phi(r_1)}{\partial x_i}, ..., \frac{\partial \Phi(r_N)}{\partial x_i})^T$. Using the gaussian basis function of equation (4) and the Euclidean norm, we can write that:

$$\frac{\partial \Phi(r_l)}{\partial x_i} = -\frac{2}{a_f^2} \delta X(i,l) \Phi(r_l) \tag{7}$$

To compute the Hessian matrix, we can do the same with the equation (6). So we have:

$$H(\tilde{f}(X))_{ij} = \frac{\partial^2 \tilde{f}(X)}{\partial x_{ij}^2} = Kij^T \cdot W \tag{8}$$

where $Kij = (\frac{\partial^2 \Phi(r_1)}{\partial x_{ij}^2}, ..., \frac{\partial^2 \Phi(r_N)}{\partial x_{ij}^2})^T$.

As we have done in (7), we can write:

$$\frac{\partial^2 \Phi(r_l)}{\partial x_{ij}^2} = \Phi(r_l)[\frac{4}{a_f^2}\delta X(i,l)\delta X(j,l) - \frac{2}{a_f^2}\delta_{ij}] \tag{9}$$

where δ_{ij} is the Kronecker symbol.

The implementation of such derivatives in an optimization program is straightforward and the computational cost is negligible when comparing it to the cost of one exact evaluation of the function f.

4 Kriging metamodel

Kriging is a spatial interpolating technique of a given function f from a set of observations $(X_i,\ f(X_i) = f_i)$. The word kriging is derivated from the name of the South African mining engineer Daniel Gerhardus Krige who proposed, in the fifties, the fundamental principle of this statistical method in order to find the spatial distribution of gold on the Witwatersrand. [8]. However, it was the french mathematician Georges Matheron who formalised the theory of this technique in the sixties and called it "krigeage". [10]. This technique is also known as gaussian process regression or best linear unbiased prediction [1].

In the literature, one can find several ways to present this technique. See for instance [9], [4], [1] or [7]. The presentation of D. R. Jones [7]("a gentle introduction to kriging") is of particular interest. In the present paper, we will introduce the kriging with two points of view: using the variance approach and using the joint probability density approach. We will see that this two ways leads to the same results when the gaussian distribution is assumed. We will be limited here to what is called simple kriging. The undermentioned introduction is not a full presentation of the kriging theory. The reader interested can refer to the above articles or to somes related references such as [14].

4.1 The kriging using the variance approach

This approach can be found for exemple in [1] and in [12]. It is general because it does not assume any particular distribution of the random variables. But the Gaussian distribution will be needed when optimizing the correlation function.

As for the other metamodel techniques, the kriging suggests a predictor $\tilde{f}$ of a function f on a point X given the values of this function on the points X_i $(i = 1, N)$. This predictor is a sort of linear combination between the known values $f_i = f(X_i)$ of the function at the observed points:

$$\tilde{f}(X) = \sum_{i=1}^{N} \omega_i(X) f_i \tag{10}$$

where $\omega_i(X)$ are some weights to be determined each time we want to evaluate the function. To simplify the notation, we will note these weights ω_i but we must keep in mind that they are functions of X.

The main idea of the kriging is to determine these weights using some statistical considerations. The kriging supposes that the value $f(X)$ of the function at any point of the design space is a realisation of a random process. This value is thus a random variable that will be noted $F(X)$ and will take the value $f(X)$ after the randon experiment. The expectation of $F(X)$ is $\mu(X)$ and its variance is $var(X)$. The kriging supposes also that the values taken by this random variable at each two points X and Y of the design space are correlated and that this correlation is known and is a given symmetric spatial function $c(X,Y)$ of the locations. That is:

$$cov[F(X),F(Y)] = c(X,Y) \tag{11}$$

where $cov[F(X),F(Y)]$ is the covariance between $F(X)$ and $F(Y)$.

If we write the equation (10) using the random variable notation we will get:

$$\tilde{F}(X) = \sum_{i=1}^{N} \omega_i F(X_i) \tag{12}$$

Thus, the predictor $\tilde{F}(X)$ is also a random variable whose values depend on the values taken by the random field $(F(X_1),...,F(X_N))$. Its expectation is $\mu[\tilde{F}(X)]$ and its variance is $var[\tilde{F}(X)]$. In the same way, the error $e(x) \equiv \tilde{F}(X) - F(X)$ of the prediction of $F(X)$ by $\tilde{F}(X)$ (kriging error) is a random variable whose values depend on those taken by the randon field $(F(X_1),...,F(X_N),F(X))$. Its expectation is $\mu[e(X)]$ and its variance is $var[e(X)]$. Starting from the equation (12) and the definition of the error, we can see that $var[\tilde{F}(X)]$ and $var[e(X)]$ are given by:

$$var[\tilde{F}(X)] = cov[\tilde{F}(X),\tilde{F}(X)] = \sum_{i=1}^{N}\sum_{j=1}^{N} \omega_i\omega_j c(X_i,X_j) \tag{13}$$

and

$$var[e(X)] = var[\tilde{F}(X)] + var(X) - 2cov[\tilde{F}(X),F(X)] \tag{14}$$

We know also that:

$$var(X) = cov(X,X) = c(X,X) \tag{15}$$

and

$$cov[\tilde{F}(X),F(X)] = \sum_{i=1}^{N} \omega_i c(X,X_i) \tag{16}$$

We can thus conclude that:

$$var[e(X)] = \sum_{i=1}^{N}\sum_{j=1}^{N} \omega_i\omega_j c(X_i,X_j) + c(X,X) - 2\sum_{i=1}^{N} \omega_i c(X,X_i) \tag{17}$$

In the same manner, we can develop a similar expression for the expectation of the error, which gives:

$$\mu[e(X)] = E[e(X)] = \sum_{i=1}^{N} \omega_i E[F(X_i)] - E[F(X)] = \sum_{i=1}^{N} \omega_i \mu(X_i) - \mu(X) \tag{18}$$

In the ideal case, i.e if the predictor is perfect, it always givses the value taken by $F(X)$. This means that the error vanishes and so $\mu[e(X)] = 0$ and $var[e(X)] = 0$ for all X. In the real case, in general, this can not be true. Hence, to get the best predictor, we must find the values of the weights ω_i so that $\tilde{F}(X)$ is the closest to the perfect predictor. This means that $\mu[e(X)]$ must be equal to zero and $var[e(X)]$ must be the smallest. The first condition is the unbiasedness condition (the mean value of $F(X)$ is equal to the mean value of $\tilde{F}(X)$) and the second maximizes the "trust" we have in the expected value. So, looking for the best predictor becomes a minimization problem of equation (17) under the constraint:

$$\sum_{i=1}^{N} \omega_i \mu(X_i) - \mu(X) = 0 \tag{19}$$

There are several kinds of kriging that can be distinguished by the assumptions used to simplify the above minimization problem. Among this kinds we cite those that are the most commonly used [1]:

- Simple kriging supposes that the expectation $\mu(X)$ is equal to zero[1], so the unbiasedness condition is satisfied whatever are the values of ω_i.
- Ordinary kriging supposes that the expectation has a fixed but unkown value μ[2]. The unbiasedness condition (19) becomes simply $\sum_{i=1}^{N} \omega_i = 1$. This constraint is taken into account by mean of a Lagrange multiplier.
- The universal kriging suppose that the expectation is a given function of X.

In this study, we are interested only in simple kriging. The expectation $\mu(X)$ is equal to zero and the predictor is unbiased. In the minimization of (17) the correlation function is assumed to be a fixed parameter. So minimizing this equation can be done easily by derivating it with respect to ω_i. Hence, for each ω_i we obtain the equation:

[1] Sometimes, simple kriging is used under the assumption that the expectation is not equal to zero but have a fixed and known value μ. In this case, the simple kriging expressions are applied to the function $g(X) = f(X) - \mu$ and the kriging predictor becomes $\tilde{f}(X) = \mu + \tilde{g}(X)$.

[2] One can think that this is equivalent to use simple kriging with a non-zero expectation. It is not the case at all! Note that the linear expression (12) is applied to $F(X)$ in the case of ordinary kriging and to $F(X) - \mu$ in the case of simple kriging. Even if, in simple kriging, μ is often given by a linear combination of f_i (which means that $\tilde{f}(X)$ is a linear combination of f_i but $\tilde{F}(X)$ is not a linear combination of $F(X_i)$) the two methods does not lead to the same expressions.

$$\sum_{j=1}^{N} \omega_j c(X_i, X_j) - c(X_i, X) = 0 \tag{20}$$

This is equivalent to the linear system:

$$C_N \cdot W = K \tag{21}$$

where C_N is the correlation matrix defined by $C_N(i,j) = c(X_i, X_j)$, $W = (\omega_1, ..., \omega_N)^T$ and $K = (c(X_1, X), ..., c(X_N, X))^T$.

Finally, from (10) and (21), the predicted value of $f(X)$ is given by:

$$\tilde{f}(X) = K^T C_N^{-1} \mathbf{F_N} \tag{22}$$

Where the vector $\mathbf{F_N}$ is given by $\mathbf{F_N} = (f_1, ..., f_N)^T$.

If we note $k = c(X, X)$, and by combining equations (17) and (21), we can easily conclude that:

$$var[e(X)] = k - K^T C_N^{-1} K \tag{23}$$

This expression is an estimation of the uncertainty of the predicted value. This is one of the advantages of kriging. It gives, not only a prediction of the unkown function value, but also an indication about the error of the metamodel.

In the equations (22) and (23), if X is one of the observed points, let say X_i, the vector K becomes the i^{th} column of the correlation matrix, and so we get that $\tilde{f}(X_i) = f_i$ and $var[e(X)] = 0$. This means that the predictor is "perfect" on the observed points; it exactly reproduces the function at this points. It is thus an interpolating model.

Note that we have not assumed any distribution of the random variable $F(X)$. A priori, this can be any distribution because the above expressions are very general. But as we will see later, the correlation function needs to be optimized to get the best predictor. To make this optimization straightforward, the Gaussian distribution is assumed.

4.2 The kriging using the joint probability density approach

This approach is perhaps the most popular presentation of kriging found in the literature. In this case, we assume that the random variable $F(X)$ (see (4.1)) is normally distributed. In the case of simple kriging, we assume also that $\mu(X) = 0$ (see also note (1)). Let $\sigma_1(X)$ be the standard deviation of $F(X)$ ($\sigma_1^2(X) = var(X)$). Then, the probability density of $F(X)$ at each point X is given by:

$$p(f(X)) = \frac{1}{\sqrt{2\pi\sigma_1(X)}} exp\left(-\frac{f(X)^2}{2\sigma_1^2(X)}\right) \tag{24}$$

The generalization of this distribution to the random fields $(F(X_1), ...F(X_N))^T$ and $(F(X_1), ...F(X_N), F(X))^T$ is done using the joint probability density formulation of a multivariate Gaussian process. This yields, for these fields respectively:

$$p(\mathbf{F_N}) = \frac{\mathbf{1}}{\sqrt{(\mathbf{2}\pi)^{\mathbf{N}}\mathbf{det(C_N)}}}\mathbf{exp}\left(-\frac{\mathbf{1}}{\mathbf{2}}\mathbf{F_N^T}\mathrm{C_N^{-1}}\mathbf{F_N}\right) \tag{25}$$

and

$$p(\mathbf{F_{N+1}}) = \frac{\mathbf{1}}{\sqrt{(\mathbf{2}\pi)^{\mathbf{N+1}}\mathbf{det(C_{N+1})}}}\mathbf{exp}\left(-\frac{\mathbf{1}}{\mathbf{2}}\mathbf{F_{N+1}^T}\mathrm{C_{N+1}^{-1}}\mathbf{F_{N+1}}\right) \tag{26}$$

where $\mathbf{F_N} = (f_1, ...f_N)^T$ and $\mathbf{F_{N+1}} = (f_1, ...f_N, f(X))^T$. These vectors are the values taken by the random fields $(F(X_1), ...F(X_N))^T$ and $(F(X_1), ...F(X_N), F(X))^T$ respectively. $det(C_N)$ represents the determinant of the matrix C_N. The matrix C_{N+1} has the same structure as C_N for which we add one line and one column as follows (see (4.1) for notations):

$$C_{N+1} = \begin{pmatrix} C_N & K \\ K^T & k \end{pmatrix} \tag{27}$$

Since we know that the field $(F(X_1), ...F(X_N))^T$ takes the values of $\mathbf{F_N}$, we are not interested in evaluating the joint probability density of equations (25) and (26) itself. All we want is the probability density of $F(X)$ knowing the observation $\mathbf{F_N}$. This is given by the conditional probabilities rule:

$$p(f(X)/\mathbf{F_N}) = \frac{\mathrm{p\,(f(X) \cap F_N)}}{\mathrm{p\,(F_N)}} = \frac{\mathrm{p(F_{N+1})}}{\mathrm{p(F_N)}} \tag{28}$$

From equations (25), (26) and (28), we can conclude that:

$$p(f(X)/\mathbf{F_N}) = \sqrt{\frac{\mathrm{det(C_N)}}{2\pi\mathrm{det(C_{N+1})}}}\exp\left[\frac{1}{2}\left(\mathbf{F_N^T}\mathrm{C_N^{-1}}\mathbf{F_N} - \mathbf{F_{N+1}^T}\mathrm{C_{N+1}^{-1}}\mathbf{F_{N+1}}\right)\right] \tag{29}$$

To evaluate this probability density, we need to find the expressions of $\frac{det(C_N)}{det(C_{N+1})}$ and $\left(\mathbf{F_N^T}\mathrm{C_N^{-1}}\mathbf{F_N} - \mathbf{F_{N+1}^T}\mathrm{C_{N+1}^{-1}}\mathbf{F_{N+1}}\right)$. This can be done by expressing C_{N+1}^{-1} in terms of linear combination of C_N^{-1} and the vector K. In this paper, we do not describe all the algebraic details leading to this expressions because this can be cumbersome. Some of these details are given in [2] [4] and [13] for instance. Here we will just give the final results:

$$\mathbf{F_N^T}\mathrm{C_N^{-1}}\mathbf{F_N} - \mathbf{F_{N+1}^T}\mathrm{C_{N+1}^{-1}}\mathbf{F_{N+1}} = -\frac{\mathbf{(f(x) - f_0)^2}}{\boldsymbol{\sigma^2}} \tag{30}$$

and

$$det(C_{N+1}) = \sigma^2 det(C_N) \tag{31}$$

where

$$f_0 = K^T C_N^{-1}\mathbf{F_N} \tag{32}$$

and

$$\sigma^2 = k - K^T C_N^{-1} K \tag{33}$$

This means that the random event $F(X)/\mathbf{F_N}$ is also a realisation of a gaussian process of mean f_0 and of standard deviation σ. Its distribution is given by substituting expressions (30) and (31) into (29). This leads to:

$$p(f(X)/\mathbf{F_N}) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(\mathrm{f(X)} - \mathrm{f_0})^2}{2\sigma^2}\right) \tag{34}$$

Using this approach, the kriging predictor is nothing else than the mean of the event $F(X)$ knowing the observation $\mathbf{F_N}$ ($\tilde{f}(X) = f_0$). Note that expressions (32) and (33) are identical to expressions (22) and (23) respectively. Hence, in simple kriging (at least), the mean of the event $F(X)/\mathbf{F_N}$ is the best unbiased linear predictor[3] in the sense of error variance minimization, and the variance of the kriging error is equal to the variance of $F(X)/\mathbf{F_N}$.

Note that expressions (25), (26) and hence (29) are not mathematically definites if $det(C_N)$ or $det(C_{N+1})$ are not strictly positives. The correlation matrix is positive definite if all points X_i are distincts each other. In this case $det(C_N)$ is strictly positif. It is also the case for $det(C_{N+1})$ if the point X is not one of the observed points X_i. If it is, $det(C_{N+1}) = 0$. So we can not evaluate expression (26), (29) nor (34). But this is not a limitation. Indeed, on the one hand, one can notice that if X is one of the observed points, let say X_i, the event $F(X_i) = f_i/\mathbf{F_N}$ is certain. Then, we do not need to calculate $p(f(X)/\mathbf{F_N})$ and we know automatically that $f_0 = f_i$. On the other hand, we have seen in (4.1) that equation (22) does exactly reproduce the observed point. So, from this remarks, we conclude that the kriging metamodel given by (22) or (32) is definite for all values of X and that the error is zero at the observed points.

4.3 The correlation function

The choice of the correlation function is a very important issue in kriging. As we can notice from equations (32), this function must reflect the behaviour of the predicted function, because the metamodel is nothing else than a linear combination of the functions $c(X, X_i)$. But in general, we do not have enough information on this behaviour. Thus, a general form of this function is very often used. An other important aspect to be taken into account in the choice of the correlation function, is the fact that if two points X and Y are too far each other, their correlation tends to a small value, and if they are very close, it tends to a finite value. Thus, the correlation function must be somewhat a function of the distance between X and Y. The most commonly used form is an exponential one, which has several mathematical properties. In this paper, the correlation function takes the form:

$$c(X,Y) = \theta_1 exp\left[-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i - y_i)^2}{r_i^2}\right] + \theta_2 \tag{35}$$

[3]The predictor is linear in the sense that it is a linear combination of f_i but is not linear in X since the weights ω_i are non-linear functions of X.

In this expression, θ_1 scales the correlation between X and Y, θ_2 gives an offset from zero and r_i scales the distance between the components x_i and y_i of the vectors X and Y [4]. These parameters need to be determined using a statistical criteria. A classical technique in statistiques to determine the parameters of a given distribution is the maximum of the likelihood function. Let $\Theta = (\theta_1, \theta_2, r_1, ..., r_n)$. The likelihood function $\mathrm{L}(\Theta)$ for the observation $\mathbf{F_N}$ is equal to the the joint probability density of this observation, calculated with the parameters Θ. That is:

$$\mathrm{L}(\Theta) = \mathrm{p}(\mathbf{F_N}) \tag{36}$$

Maximizing the likelihood function is finding the best vector of parameters Θ that maximizes the probability that the values taken by the field $(F(X_1), ...F(X_N))^T$ correspond to the observation $\mathbf{F_N}$. For this purpose, a Gaussian distribution is assumed. With this assumption, it is easier to work with the log-likelihood function $\mathcal{L}(\Theta)$ defined by $\mathcal{L}(\Theta) = -2\mathrm{Log}(\mathrm{L}(\Theta))$ instead of L itself (the factor 2 is here only for simplifying the expression and it does not change the result). Indeed, from equations (25) and (36) we can write, when omitting the constant term $N\mathrm{Log}(2\pi)$:

$$\mathcal{L}(\Theta) = \mathrm{Log}\left(\det(\mathrm{C_N})\right) + \mathbf{F_N^T} \mathrm{C_N^{-1}} \mathbf{F_N} \tag{37}$$

Hence, maximizing the likelihood function is equivalent to minimizing the function $\mathcal{L}(\Theta)$. Here again, we have an optimization problem which is multimodal. As for the Radial Basis Functions metamodel, this optimization problem can be solved using the PSO method to avoid local minima. To evaluate the cost function (37), the determinent of the correlation matrix can be computed, for instance, using the LU decomposition.

4.4 Differentiation of the kriging metamodel

As for the RBF metamodel, It is easy to evaluate the gradient and the Hessian of the predicted function and with low cost. This is made possible by the choice of the correlation function of the form of (35) which is a smooth function. The differentiation of the kriging metamodel is very similar to that of the RBF metamodel. So, for the gradient we have:

$$grad(\tilde{f}(X))_i = \frac{\partial \tilde{f}(X)}{\partial x_i} = Ki^T \cdot C_N^{-1} \mathbf{F_N} \tag{38}$$

where $Ki = (\frac{\partial c(X,X_1)}{\partial x_i}, ..., \frac{\partial c(X,X_N)}{\partial x_i})^T$.

Let $h(X,Y) = \theta_1 exp\left[-\frac{1}{2}\sum_{i=1}^{n} \frac{(x_i - y_i)^2}{r_i^2}\right]$ and $\delta X = (X - X_1, ..., X - X_N)$. Using the expression of the correlation function (35), we can write that:

$$\frac{\partial c(X, X_l)}{\partial x_i} = -\frac{1}{r_i^2} \delta X(i,l) h(X, X_l) \tag{39}$$

for $l = 1, N$.

To compute the Hessian matrix, we can apply the same process with the equation (38). So we have:

$$H(\tilde{f}(X))_{ij} = \frac{\partial^2 \tilde{f}(X)}{\partial x_{ij}^2} = Kij^T \cdot C_N^{-1}\mathbf{F_N} \tag{40}$$

where $Kij = (\frac{\partial^2 c(X,X_1)}{\partial x_{ij}^2}, ..., \frac{\partial^2 c(X,X_N)}{\partial x_{ij}^2})^T$.

As we have done in (39), we can write:

$$\frac{\partial^2 c(X,X_l)}{\partial x_{ij}^2} = h(X,X_l)\left[\frac{1}{(r_i r_j)^2}\delta X(i,l)\delta X(j,l) - \frac{1}{r_i^2}\delta_{ij}\right] \tag{41}$$

for $l = 1, N$ (δ_{ij} is the Kronecker symbol).

5 Comparaison between RBF and Kriging metamodels

Kriging and RBF metamodels are both interpolation techniques that give a prediction of the value of a function f at any point X. The philosophies of these metamodels are very different. The kriging uses a statistical approach to predict the function, whereas the RBF uses a classical interpolation approach based on the distance from the known function value points. Even though, there are a lot of similarities between the final expressions of their predictors. Under some assumptions, these predictors can have the same expressions.

When we examine expressions (2) and (3), the former can be rewritten as follows:

$$\tilde{f}(X) = K_{RBF}^T A^{-1}\mathbf{F_N} \tag{42}$$

Where $K_{RBF} = (\Phi(X,X_1), ..., \Phi(X,X_N))^T$ and $\mathbf{F_N} = (f_1, ...f_N, f(X))^T$ is the vector of observations. This is very similar to the expression of the simple kriging predictor (22) in which we have remplaced the vector K by K_{RBF} and the matrix C_N by A. Let K_{krig} be the vector K in (22). When we look at K_{RBF} and K_{krig}, we notice that they have the same structure, where the function $\Phi(X,X_i)$ in K_{RBF} is remplaced by $c(X,X_i)$ in K_{krig}. This is also the case for matrices A and C_N where $\Phi(X_i,X_j)$ in A is remplaced by $c(X_i,X_j)$ in C_N.

Suppose now that the RBF predictor uses the Gaussian function of (4), that the distance is calculated using the Euclidean norm, and that the correlation function of the kriging pridictor takes the form of (35). By examining these functions, one can notice that the radial basis function $\phi(\parallel X - X_i \parallel)$ is nothing else the the correlation function $c(X,X_i)$ where we set the parameters vector Θ to $(1, 0, a_f, ...a_f)$. Hence the two predictors are exactly the same in this case!

The use of the parameters r_i in the kriging correlation function is a way to find the best scale for each component of the design vector. This is an advantage of kriging but, by itself, is not a fundamental difference between the two metamodels. The main difference is the way the metamodel parameters (a_f or Θ) are determined. In the RBF metamodel the attenuation factor is obtained by minimizing the leave-one-out error, whereas in kriging, the correlations parameters are obtained by maximizing the likelyhood function. These are very

different criteria and they may lead to different results. It is not easy to state whether there is a relationship between these criteria or not and which of them is better. This may be an interesting issue.

6 Difficulties when using metamodels

In general, the matrices involved in the RBF and kriging metamodels are ill-conditioned. In practice, this phenomenon has two impacts: the limitation of the number of the data points (the condition number increases with the size of the matrix) and the loss of the accuracy of the model. Indeed, when the condition number is high, the accuracy of the matrix inversion is very bad. So, to limitate these impacts, it is recommended to use an adequate algorithm to compute the matrix inverse.

To get an accurate model, one must well represent the function by the data set. This means that the observed points must be spread out onto the entire design space and their number must be large enough to get information about the function to be predicted. The location of these points is also a relevent issue. For oscillating functions for instance, a large number of data points is required. This number increases with the dimension n of the design vector. Suppose that we would like to construct a grid in which we subdivide the design space into four sections only on each directions. The number of points of this grid is 16 if $n = 2$ and 65536 if $n = 8$. This exemple highlights the effect the dimension of the design vector on the required number of data points. With a large number of data points, the training of the metamodel becomes stiff, if possible. Thus, for high dimension design vectors, it is difficult to get enough data points. It would then be better to use a local metamodel i.e a metamodel constructed with a few number of data arroud the point where the function needs to be predicted [2]. This is a limiting technique because it requires a training of the model for each evaluation and it supposes enough available data arround each point. An other limiting problem with the number of data set is their generating cost. One must keep in mind that the main advantage of a metamodel is the reduction of the evaluation cost of the desired function. This cost must not be increased by the increase of the number of data and hence by their generation cost.

In optimization problems, to avoid this difficulty, the metamodel may be constructed with a limited data size in a first stage. This permits to get a coarse information on the location of the optimum. Then the metamodel is updated by adding new points arroud this optimum in the data set, and the routine is ran again until localising the optimum [3] [7] [5].

7 Conclusion

In this paper we present a review of the Radial Basis Functions and the kriging metamodels which are widely used in optimization and other applications. The formulation of these techniques is described and their differentiation is given. The kriging metamodel is presented

using the variance and the joint density probability approaches. It is shown that these two different approches lead to the same formulation of the kriging predictor. In addition, the comparison between the RBF and the kriging metamodels shows that they are very similar. Under some assumptions, these two metamodels can have the same formulation. The main difference is the way the parameters are trained. The former uses, for instance, the leave-one-out technique whereas the later uses the likelihood maximization one.

In this paper we discuss also the difficulties related to the construction of these metamodels. One of these difficulties is the location of the observation points. Indeed, if the predicted function presents many variations, the observation points must be chosen such that the information obtained allows to reproduce the behaviour of this function. This means also a large number of the observation points. But the most important difficulty is the size of the data set. When this size is too large, it becomes too expensive to construct the metamodel. To avoid this difficulty, some authors propose to construct a local metamodel, with only few observation points arround the desired one. The resulting metamodel is then cheaper. But this requires enough available data in the vicinity of each predicted point.

8 Acknowledgments

This study hase been supported by the "OMD" project (Multi-Disciplinary Optimization) granted by ANR-RNTL.

References

[1] Kriging. In *Wikipedia, The Free Encyclopedia*, http://en.wikipedia.org/wiki/Kriging, (accessed: 2009, February, 13).

[2] D. Büche, N. N. Schraudolph, and P. Koumoutsakos. Accelerating evolutionary algorithms with gaussian process fitness function models. *IEE Transactions on Systems, Man and Cybernetics - Part C: Applications and Reviews*, 35 (2):183–194, 2005.

[3] P. Chandrasherakarappa and R. Duvigneau. Metamodel-assisted particle swarm optimization and application to aerodynamique shape optimization. *Resarch Report, INRIA*, 6397, 2007.

[4] P. Chandrasherakarappa and R. Duvigneau. Radial basis functions and kriging metamodels for aerodynamic optimization. *Resarch Report, INRIA*, 6151, 2007.

[5] K. C. Giannakoglou and M. K. Karakasis. Hierarchical and distributed metamodel-assisted evolutionary algorithms. In *Introduction to Optimization and Multidisciplinary Design*. von Karman Institue of Fluid Dynamics, Lecture Series, March 2006.

[6] Y. Jin. A comprehensive survey of fitness approximation in evolutionary computation. *Soft Computing*, 9 (1), 2005.

[7] D. R. Jones. A taxonomy of global optimization methods based on response surfaces. *J. Global Optimization*, 21:345–383, 2001.

[8] D. G. Krige. A statistical approach to some basic mine valuation problems on the witwatersrand. *J. of the Chem., Metal. and Mining Soc. of South Africa*, 52 (6):119–139, 1951.

[9] J. D. Martin and T. W. Simpson. Use of kriging models to approximate deterministic computer models. *AIAA Journal*, 43 (4):853–863, 2005.

[10] G. Matheron. *Traité de géostatistique appliquée*. Technip, 1962.

[11] S. Rippa. An algorithm for selecting a good value for the parameter c in radial basis function interpolation. *Adv. Comp. Math.*, 11:193–210, 1999.

[12] J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn. Design and analysis of computer experiments (with discussion). *Statistical Science*, 4:409–435, 1989.

[13] H. Theil. *Principles of Econometrics*. John Wiley, New York, 1971.

[14] H. Wackernagel. *Multivariate Geostatistics - An Introduction with Applications*. Springer, Berlin, 1995.

Contents

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Unité de recherche INRIA Futurs : Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)
Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399